Journal of Computing and Information Technology - CIT 18, 2010, 4, 325–332
doi:10.2498/cit.1001912

Contextualized Question Answering

Luka Bradeško, Lorand Dali, Blaž Fortuna, Marko Grobelnik, Dunja Mladenić, Inna Novalija and Boštjan Pajntar

Jožef Stefan Institute, Ljubljana, Slovenia

The paper describes a system which enables accurate and easy-to-use contextualized question answering and it provides document overview functionalities. The possibility of asking natural language questions enables a friendly interaction for the user.The contextualization is achieved by using an ontology. The answers are provided based on a domain specific document collection of choice. The approach consists of several phases as follows: data preparation, data enhancement, data indexing and handling questions. Every module uses state of the art technologies that are shown to work in a complex pipeline to make available question answering on top of a given document repository with the context of ontologies, such as Cyc, ASFA and WordNet. The functioning of the proposed approach is demonstrated on English document collections on Aquatic Sciences and Fisheries — ASFA, using Cyc ontology, ASFA thesaurus as domain specific ontology and WordNet as general ontology. Experimental evaluation has shown that the usage of ontologies increases the number of answers retrieved by about 60%. However, the number of answers that are actually correct increases by only 40% when using ontologies.

Keywords: contextualized information retrieval, question answering, ontology

1. Introduction

Search applications are very useful and widely used. However, most of them only retrieve documents (or snippets) which match the search query best, without a specific answer or any information related to the context. Our contribution is a search application which enables querying in natural language and provides specific answers to the question asked, moreover, domain specific and general context is leveraged to retrieve facts which are not explicitly stated. The application also provides document overview functionality which enables the user to explore further information in the context of the current answer.

The technology behind is based on AnswerArt[4] technology for question answering and Cyc[7] ontology for providing semantic context to the document collection from a particular domain of interest.

The approach consists of several phases as follows. In the **data preparation phase**, we extract the relevant part of Cyc ontology and extend it with any other relevant ontology, either general or domain specific. In this phase, the document collection is pre-processed using AnswerArt technology to obtain subject-predicate-object triplets from the sentences in the domain specific document collection. In the **data enhancement phase**, the triplets are enhanced using semantic knowledge obtained from the extended part of Cyc ontology. In the **data indexing phase**, the enhanced triplets are index for efficient search for answers. In the final phase of **handling question**, the question is transformed based on predefined patterns of questions to enable efficient search over the indexed triplets and the list of answers is returned.

The remainder of this paper is structured as follows. Section 2 provides a brief description of the motivation behind this work. Section 3 describes the related work, Section 4 describes the approach giving an overview of the architecture. In Section 5 we briefly describe the underlying technology. Section 6 describes the ASFA data used for answering questions, while the details on the extension of Cyc ontology using WordNet[3, 5] and ASFA ontology are in Section 7. Usage of the system is illustrated in Section 8. Section 9 describes the evaluation and Section 10 concludes with a short discussion and ideas for future development.

2. Motivation

When performing search, users are frequently looking for specific answers to questions. Moreover, we consider additional facts from the context of the answer, and document overview functionality as being helpful for the user. We also think that enabling natural language queries makes the system much more user friendly.

The proposed approach is based on combining AnswerArt [4] technology for question answering based on triplets and Cyc ontology for providing semantic context to the document collection from a particular domain of interest. Moreover, the approach provides answers in semantic context of the domain of interest via including Cyc ontology extended by any other domain specific ontology as needed.

AnswerArt integrates two important functionalities: providing answers to questions and browsing through documents that support the answers. The questions follow a predetermined template, whereas the answers are yielded based on the previously extracted information, in the form of subject–predicate–object triplets. Furthermore, the system retrieves the sentences that support these answers, as well as the documents containing the sentences. It integrates three possibilities of further exploring the relevant documents: by analyzing the list of facts extracted from the document, by visualizing the semantic representation of the document and by browsing the document summary.

3. Related Work

Related approaches query structured data stored in ontologies, while AnswerArt derives the answers only from unstructured text. TextRunner [1] is similar to AnswerArt in the way that it also involves applying structured queries on unstructured text, while the main difference is that AnswerArt also provides a natural language interface to the search. The Calais[1] system creates semantic metadata for user submitted documents in the form of named entities, facts and events. AnswerArt named entities and facts represent the starting point and they are further refined by applying co-reference resolution for named entities, anaphora resolution and semantic normalization based on WordNet for facts. This process enables the construction of a semantic description of the document in the form of a semantic directed graph where the nodes are the subject and object triplet elements, and the link between them is determined by the predicate. Powerset[2] enables search over Wikipedia and Freebase, where the search results contain aggregated information from several articles, as well as a list of facts related to people, places and things. The main difference is that AnswerArt describe the answer by a visual representation of the document in the form of a semantic graph and by the document summary, which is automatically extracted based on the document semantic graph.

4. Approach Description

We propose an approach that enables answering question from a desired domain using a collection of relevant documents. The approach consists of four phases: data preparation, data enhancement, data indexing and question handling. Architecture of the proposed approach is given in Figure 1.

In the data preparation phase, we extract the relevant part of Cyc ontology and extend it with any other relevant ontology, either general or domain specific. In our application scenario, we have selected ASFA abstracts for a document collection and extended Cyc by using WordNet and ASFA ontology. The document collection is pre-processed using AnswerArt technology to obtain subject-predicate-object triplets out of the sentences in the document collection.

In the data enhancement phase, the triplets are enhanced using semantic knowledge obtained from the extended Cyc ontology. In particular, each part of the triplet is extended by a set of synonyms and direct generalizations obtained from the ontology.

In the data indexing phase, the extended triplets are indexed for efficient search for answers. The search is performed by transforming each question into a set of semi-triplets — triplets with missing one or two arguments. Search is performed as matching of the semi-triplets with the triplets from the index in order to find possible

[1] Calais web page: `http://www.opencalais.com/`
[2] Powerset web page: `http://www.powerset.com/`

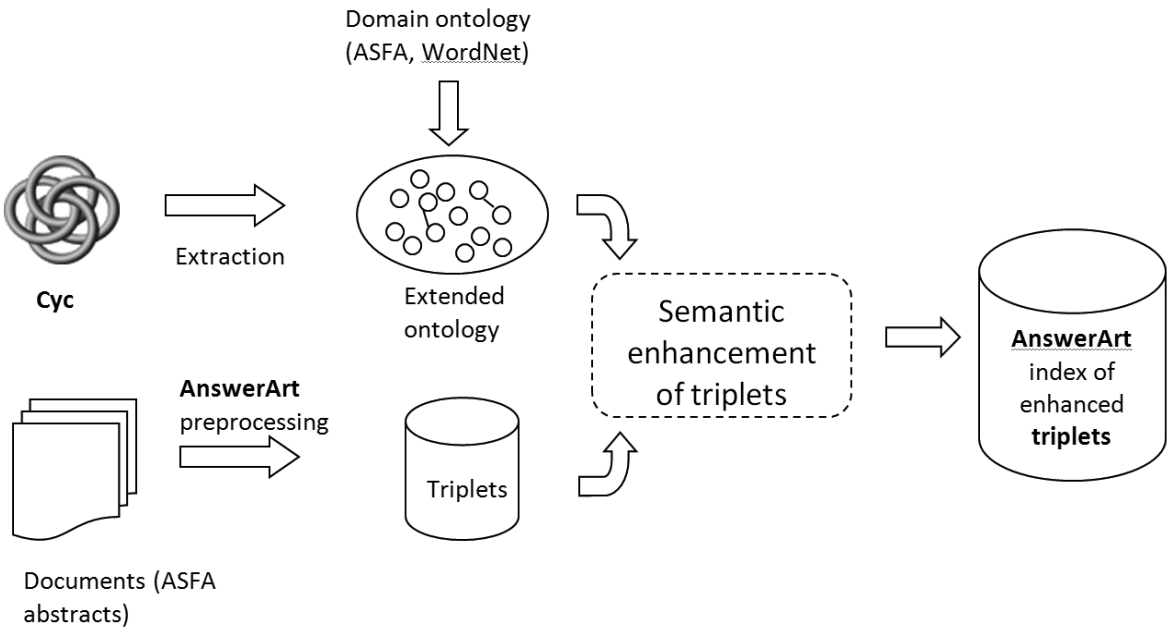

Figure 1. Architecture of the approach.

values of the missing arguments (answers to the question).

In the final phase, the question is transformed based on predefined patterns of questions to enable efficient search over the indexed triplets and the list of answers is returned. In the question handling phase, the question is transformed based on predefined patterns of questions to enable efficient search over the indexed triplets and the list of answers is returned.

5. Underlying Technology

The proposed approach builds on several existing technologies: components of the AnswerArt system for question answering are adjusted for including semantic enhancement of the data using Cyc ontology. It is implemented in a prototype that uses service oriented architecture. It consists of module for the triplet extraction, which can work over any document repository. In our implementation, we demonstrate it over the ASFA abstracts. Next, there is a module which consolidates Cyc and domain ontologies, which all provide semantic enhancements of the triple set. Last, there is AnswerArt module for translating natural language questions into triplet queries and visualization and summarization of the results.

5.1. AnswerArt description

AnswerArt combines question answering, summarization and document visualization functionalities. The user obtains answers based on the facts previously extracted from text in the form of subject–predicate–object triplets. Moreover, the sentences that support the answer, as well as the documents containing these sentences, are also retrieved. The relevant documents can be further explored with the aid of a document overview functionality that consists of a document summary, a semantic representation of the initial document and a list of facts extracted from the document (see Figure 2).

The system shows possible answers to the question, links them to the supporting sentences and corresponding document. The system provides a document overview by retrieving the document semantic graph, the list of subject–predicate–object facts and the automatically generated document summary of variable length interactively set by the user. Extraction of subject–predicate–object facts is a pre-processing step to document collection. Triplets are extracted from each sentence in turn. This means that to extract triplets from a document, the text in that document has to be split into sentences.

Moreover, each sentence is tokenized and the tokens are tagged with their parts of speech. After this, chunking is performed. Chunking

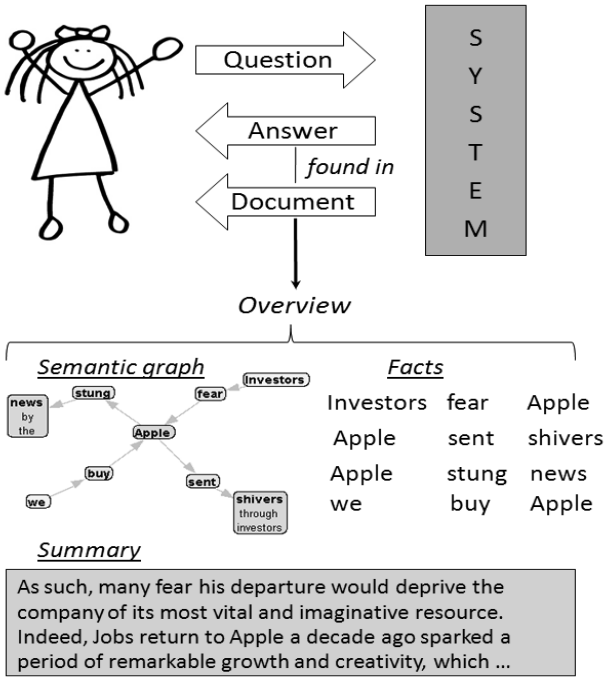

Figure 2. Illustration of AnswerArt.

means that several related consecutive tokens are grouped together, based on their tags, resulting in phrases (noun phrases and verb phrases), also called chunks. Having chunked a sentence, simple rules can be applied to extract triplets from it. An example of such a rule would be: a noun phrase followed by a verb phrase followed by another noun phrase is a triplet.

5.2. Cyc description

Cyc Knowledge Server is a very large, multi-contextual knowledge base and inference engine, developed for more than twenty years with a goal to break the "software brittleness bottleneck" once and for all by constructing a foundation of basic "common sense" knowledge — a semantic substratum of terms, rules, and relations. This enables a variety of knowledge-intensive products and services. Cyc is intended to provide a "deep" layer of understanding that can be used by other programs to make them more flexible. Here we have used a part of Cyc functionality that includes the Knowledge base, Inference Engine and The Natural Language Processing Subsystem, which we accessed through Cyc Developer Toolsets (CYC API).

The Cyc knowledge base (KB) is a formalized representation of a vast quantity of fundamental human knowledge: facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life. The medium of representation is the formal language CycL, described below. The KB consists of terms — which constitute the vocabulary of CycL — and assertions which relate those terms. These assertions include both simple ground assertions and rules.

At the present time, the Cyc KB contains nearly two hundred thousand terms and several dozen hand-entered assertions about/involving each term. New assertions are continually added to the KB by human knowledge enterers and lately also with the help of the machines using various machine learning algorithms.

6. Data Description — ASFA

Aquatic Sciences and Fisheries Abstracts (ASFA) [8] is a database covering the literature on the science, technology, management, and conservation of marine, brackish water, and freshwater resources and environments, including their socio-economic and legal aspects.

More than 5,000 serial publications, books, reports, conference proceedings, translations and limited distribution literature are selected for abstracting and indexing in ASFA. Publications in more than 40 languages, with English as primary language, are represented in ASFA database that actually aggregates five databases.

We have pre-processed the ASFA documents using AnswerArt technology. There were 3,100,832 triplets extracted from the data containing 347 403 unique terms. Examples of the extracted triplets are the following:

Salmon	spawn	rivers
Salmon	is	export product
disease	affecting	tissues
disease	cured	antibiotics
symptoms	associated	disease
weight	caught	survey
ship	deform	ice
ship	has	ballast tanks
fisheries	conservation	areas
blood	developing	mechanism
sunlight	supported	growth

7. Data Enhancement

7.1. Data enhancement with Cyc

The input for this part of the task were the subject–predicate–object triplets extracted from ASFA documents. The triplet source was then connected into the Cyc KB using OpenCYC API. For each from the vast amount of concepts (347,404) in English the Cyc KB was queried to get the corresponding Cyc concept (see the examples below). Furthermore, for each concept that Cyc has its English representation we queried it few times to get the related concepts, especially its generalizations. For each of the concepts we got one or more of its generalized meanings and then the Cyc was queried again to get the English presentations of all related concepts.

Out of the 347403 concepts occurring in the extracted triplets, 10310 are covered by Cyc. There are 228266 inferences made by Cyc, which means that one concept has roughly about 20 related concepts extracted from Cyc.

7.2. Data enhancement using WordNet

WordNet is a lexical database of the English language containing about 150 000 words organized into synsets — a set of words which have the same meaning, and also contains an explanation, examples, and the part of speech which the words in that synset have. A synset can have semantic relations to other synsets. Two of these relations relevant for our work are the hypernymy relation (more general forms) and the hyponymy relations (more specific). The WordNet database is organized into a hierarchy of hypernyms and hyponyms.

The system uses WordNet to find related terms for each concept extracted from the ASFA abstracts. By related term we mean: synonyms (the other words in the synset) and one level of hypernyms. The goal of the related terms is to improve the recall of the search for a given concept, which will be found not only if the user searches for the word as it appeared in the text, but also if he searches for a related term. There were 347404 terms extracted from ASFA, for 27775 of which synonyms could be found in WordNet, and for 20358 hypernyms could be found. For instance, catfish synonyms: mudcat; hypernyms: freshwater fish.

7.3. Data enhancement using ASFA

ASFA thesaurus contains over 9800 concepts, applies to database indexing and provides a set of terms used by indexers to describe the contents of publications. These thesaurus terms are listed in the Descriptors field of each record in ASFA database. Each concept might contain the information about hierarchical and affinitive relations with other thesaurus concepts. 4948 ASFA thesaurus concepts corresponding to 347403 ASFA abstract terms were extracted and elaborated with synonymic and hierarchical relations.

8. Illustrative Example of the System Usage

Illustrative example in Figure 3 poses a question: “What could pollution have affected”. The blue rectangle marks the input field with the question asked.

The red rectangle marks the header of the table of results, showing how the natural language question has been transformed into a triples query. The first word designates the subject, the second verb and the last the object. In our example the subject is “pollution” the verb is “affected” and we are searching for all possible objects: “the following”. In general, it is possible to have missing any single element or any two elements from the triple. It is also possible to have all three elements defined and the question actually checks weather the question is true — or more accurately — if such a claim is found in the document repository. The yellow rectangle marks the table of results, where each row is a found triple that satisfies the triple query given in the header. The green rectangle marks the first result — excerpt from the actual document that the triple was extracted from. At the top, the title is shown and clicking leads to the detailed view of the specific document.

9. Evaluation

To evaluate the contribution of the triplet enhancement with ontologies to the performance of the question answering, we conducted the following experiment. We asked 27 questions

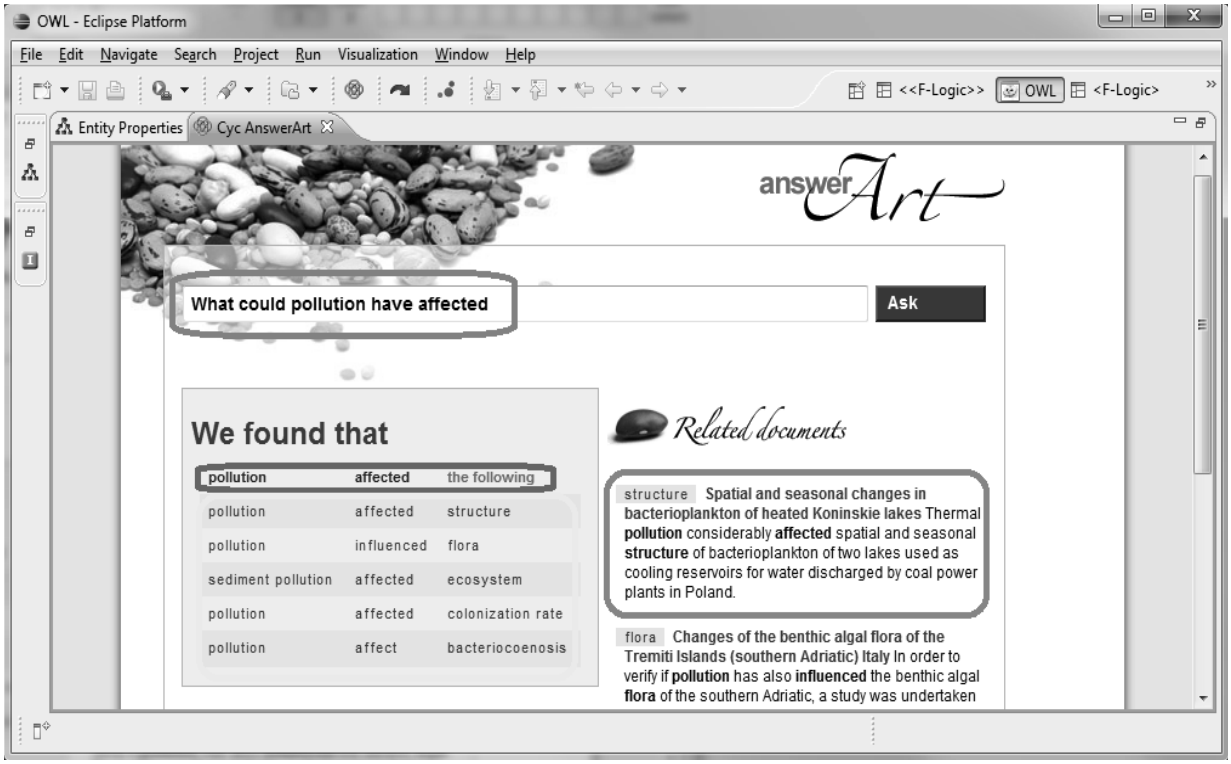

Figure 3. Example of the system usage.

to which the system responded both with and without using inference from ontologies. To each question the system gave a number of answers. The correctness or relevance of the given answers was determined according to the judgement of the authors.

On average, a question was answered with 8 answers out of which on average 3 resulted from using ontologies. Hence the usage of ontologies increases the number of answers retrieved by about 60%. However, the number of answers that are actually correct increases by only 40% when using ontologies. This shows that the precision of answers obtained using ontologies is lower and that trying to obtain more answers by inference has a negative effect on the precision. Indeed, the precision of the system drops from 84.17% to 76.61% when adding answers obtained from ontologies, because the answers using ontologies have the precision of only 63.29%

Although the size of the experiment is too small to base any solid conclusions on it, we can argue that the AnswerArt system cannot find an important number of correct answers unless it uses ontologies. On the negative side however, ontologies introduce more mistakes and decrease the precision of the system.

10. Discussion

We have presented a search application which enables querying in natural language and provides specific answers to the question asked, moreover, domain specific and general context is leveraged to retrieve facts which are not explicitly stated. The application also provides document overview functionality which enables the user to explore further information in the context of the current answer.

We used a number of state of the art technologies which were shown to work in a complex pipeline to provide the described functionality.

Experimental evaluation has shown that the usage of ontologies increases the number of answers retrieved by about 60%. However, the number of answers that are actually correct increases by only 40% when using ontologies.

Future work will be focused on improving the transformation of natural language questions into triplet queries and on improving the triplet extraction component by using different text mining tools (parser, stemmer, lemmatizer) in order to boost the accuracy of triplet extraction.

11. Acknowledgments

This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under NeOn Lifecycle Support for Networked Ontologies (IST-4-027595-IP) and PASCAL2 Network of Excellence (ICT-NoE-2008).

References

[1] M. BANKO, O. ETZIONI, The Tradeoffs Between Open and Traditional Relation Extraction. In *Proc. of the ACL*, (2008) Columbus, Ohio.

[2] D. BAXTER, B. KLIMT, M. GROBELNIK, D. SCHNEIDER, M. WITBROCK, D. MLADENIĆ, Capturing Document Semantics for Ontology Generation and Document Summarization. In *Semantic knowledge management: integrating ontology management, knowledge discovery, and human language technology*, (2009) Berlin, Heidelberg: Springer, pp. 141–154.

[3] CHRISTIANE FELLBAUM, *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[4] L. DALI, D. RUSU, B. FORTUNA, D. MLADENIĆ, M. GROBELNIK, Question Answering Based on Semantic Graphs. In *Proc. of the WWW-2009 Workshop on Semantic Search (SemSearch2009)*.

[5] GEORGE A. MILLER, WordNet: A Lexical Database for English. *Communications of the ACM*, Vol. 38 (1995), No. 11, pp. 39–41.

[6] M. GROBELNIK, J. BRANK, B. FORTUNA, I. MOZETIČ, Contextualizing Ontologies with OntoLight: A Pragmatic Approach. *Informatica Journal*, Janauray 2009.

[7] D. B. LENAT, Cyc: A Large-scale Investment in Knowledge Infrastructure. *Comm. of the ACM*, Vol. 38, No. 11, November 1995.

[8] E. FAGETTI D. W. PRIVETT, J. R. L. SEARS, *Aquatic Sciences and Fisheries Thesaurus*. Descriptors used in the Aquatic Sciences and Fisheries Information System, Food and Agriculture Organization of the UN, 2009.

Received: June, 2010
Accepted: November, 2010

Contact addresses:
Luka Bradeško
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: luka.bradesko@ijs.si

Lorand Dali
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: lorand.dali@ijs.si

Blaž Fortuna
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: blaz.fortuna@ijs.si

Marko Grobelnik
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: marko.grobelnik@ijs.si

Dunja Mladenić
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: dunja.mladenic@ijs.si

Inna Novalija
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: inna.novalija@ijs.si

Boštjan Pajntar
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: bostjan.pajntar@ijs.si

LUKA BRADEŠKO is a researcher and a PhD student at Artificial Intelligence Laboratory of the Jožef Stefan Institute, working in the area of large scale information extraction and ontology building. He has been involved in several European projects including SEKT, NEON and is principal software developer for Cycorp Europe on EU project LarKC. His research interests are in text-mining and reasoning methods. Luka is experienced software engineer and has been working for several years in several software companies in Slovenia, including Hermes Plus, Epilog, Logina. His specialties are in developing java applications and integrated systems for automatic warehousing and logistics.

LORAND DALI is a computer engineer doing research in the fields of text mining, information retrieval, information visualization, natural language processing and semantic web. Lorand graduated from the Technical University of Cluj Napoca in 2008. Currently he is enrolled in the PhD program at the Jožef Stefan International Postgraduate School and works at the Artificial Intelligence Laboratory of the Jožef Stefan Institute. Lorand is developing software for question answering, personalized information retrieval, semantic annotations for text and dynamic user interfaces for data visualization.

BLAŽ FORTUNA is a senior research assistant and a PhD student at Jožef Stefan Institute in the area of kernel methods, statistical learning and semantic web with strong focus on text analysis. In the recent years he had several publications at international conferences and developed several software modules for scalable machine learning, cross-lingual information retrieval and classification, ontology learning and active learning which are part of Text Garden software environment. Blaž was at internships in Microsoft Research and in Bloomberg. He is a co-inventor on a patent on root cause analysis developed within several projects with British Telecom.

MARKO GROBELNIK is an expert in analysis of large amounts of complex data for extracting useful knowledge. In particular, the areas of expertise comprise: data mining, text mining, information extraction, link analysis, and data visualization as well as more integrative areas such as semantic Web, knowledge management and artificial intelligence. Apart from research on theoretical aspects of unconventional data analysis techniques, he has valuable experience in the field of practical applications and development of business solutions based on the innovative technologies. Marko was employed as a researcher first, at the Computer Science Department at the University of Ljubljana and later at the Department of Knowledge Technologies at Jozef Stefan Institute, Ljubljana, Slovenia, the main national research institute for natural sciences in the country. His primary focus of research and applications is intelligent data analysis which deals with unconventional scenarios going beyond classical statistical approaches and solving problems including unstructured or semi structured data. His main achievements are from the field of text-mining (analysis of large amounts of textual data), having leading role on scientific and applicative projects funded by European Commission, having projects with industries such as Microsoft Research, British Telecom, New York Times, Siemens, and organizing several international events on the related topics.

DUNJA MLADENIĆ is an expert on study and development of machine learning, data/text mining, semantic technology techniques and their application on real-world problems. She has been associated with the J. Stefan Institute since 1987, first as a student and since 1992 employed as a researcher. She is leading Artificial Intelligence Laboratory of the Jožef Stefan Institute since 2011. She got her MSc and PhD degrees in Computer Science from the University of Ljubljana in 1995 and 1998 respectively. She was a visiting researcher at School of Computer Science, Carnegie Mellon University, USA in 1996–1997 and in 2000–2001. She has published papers in refereed conferences and journals, served in the program committee of international conferences and organized international events in the area of text mining, link analysis and data mining. She is co-editor of several books including "Data Mining and Decision Support: Integration and Collaboration", Kluwer Academic Publishers 2003, "Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies" Springer 2008, "Web Mining: from Web to Semantic Web", Springer 2004, "Semantics, Web and Mining" Springer 2006, "From Web to social Web: discovering and deploying user and content profiles", Springer 2007, "Knowledge Discovery Enhanced with Semantic and Social Information", Springer 2009.

INNA NOVALIJA holds a MS degree in intellectual systems of decision taking (Ukraine, 2005) and MA degree in economics (Hungary, 2007). Currently she is working in the area of knowledge technologies and is enrolled in the PhD program at Jozef Stefan International Postgraduate School (Slovenia). Her research interests include text mining, ontologies and reasoning, language technologies, economic and business data analysis. She participated in conferences on data mining & data warehouses and information technology interfaces.

BOŠTJAN PAJNTAR is an expert in complex data visualizations featuring graph drawing, trend diagrams and semantic landscapes, data mining on text, images and graphs. All of this expertise are usually drawn together to work in different scenarios of semantic Web, information retrieval and artificial intelligence. Boštjan is employed in Artificial Intelligence Laboratory at the Jožef Stefan Institute. Earlier, he was employed in a very successful Slovenian start-up Sonce.net, which has evolved into a leading digital marketing company in Slovenia. He has been involved with several European projects including IST-World — contributing several interactive visual tools for discovery of interesting patterns and cliques in the IST community, IMAGINATION — providing technology for contextualized search on images, NeOn — developing tools for visualisation and management of ontologies and alignments, ACTIVE — developing SearchPoint solution that provides for a context-based search.